
When Safeguards Stop at the Border

Auditing How OpenAI and Anthropic Allocate Privacy Protections Across Latin American Jurisdictions¹

Miguel Andrés	Fernanda
Escobar Palta	Stephanie Rokha
<i>University of</i>	<i>Sánchez-Umaña</i>
<i>Valparaíso</i>	<i>Internet Society</i>

With
Apart Research

Abstract

Generative AI services operate across regions with uneven regulatory capacity, and how privacy safeguards are distributed across users in different jurisdictions remains underexamined. This paper develops and tests an auditable two-layer instrument that compares legal requirements in Chile and Brazil to providers' stated commitments. Layer one applies a fourteen-requirement rubric drawn from the European Union's AI Act and the General Data Protection Regulation (GDPR), validated using expert legal ground truth for Brazil's Lei Geral de Proteção de Dados (LGPD) and Chile's Law 21.719. Without being fine-tuned, the first layer matches expert annotation in 19 of 28 cells, or 67.9% agreement, a level comparable to recent compliance-classification benchmarks. Layer two applies the same retrieval-grounded judge to the privacy policies of OpenAI and Anthropic across the EU/EEA, Brazil, and Chile, anchoring each verdict to a literal evidence snippet. Among the requirements that yield informative variation, OpenAI places Brazil and Chile together in a shared "Rest of World" tier, and Anthropic gives Brazil a dedicated LGPD annex yet leaves Chile under a general non-EU regime. These findings reveal a jurisdictional discrepancy in the privacy protections the two providers publicly commit to for different users. They suggest that provider localization tracks operational institutional capacity, especially active enforcement authority, even where declared safeguards stay inconsistent across borders. This evidence is qualitative and rests on an auditable textual basis, since every inference follows from the limited variation that the six probed requirements allowed.

¹ Research conducted at the [The Global South AI Safety Hackathon](#), June 2026

1. Introduction

Large AI providers deliver the same services into jurisdictions where institutional capacity and the ability to enforce compliance vary significantly from each other, even when legal texts converge, which raises a governance question with direct consequences for users, when a system processes personal data, drives an automated decision, or moves data across a border, do people outside the European Union receive safeguards meaningfully comparable to those inside it?

We study a deployment-side failure of equal protection, in which a single provider runs one underlying model and still hands a detailed privacy and governance package to people in a jurisdiction with credible enforcement while leaving everyone else with a thinner and more generic one. For Latin American users that difference is concrete, since it decides whether transfer guarantees are ever named, whether automated decisions are framed in terms a person can inspect, and whether any practical recourse is made legible in the policy itself. Those stakes climb sharply where such systems mediate access to credit, employment, health, and other essential services, settings in which an unexplained adverse decision carries immediate and material cost, and it is precisely there that the safety frame belongs, because the injury lives in the distance between what a law promises and what a user can actually obtain.

A brief legal baseline situates the comparison, the GDPR sets out transparency, purpose limitation, data minimization, security, accountability, safeguards for cross-border transfers, and rights over automated decision-making, and the EU AI Act builds on that foundation with obligations for high-risk systems, a set of prohibited practices, and transparency duties that reach synthetic content and automated systems with significant effects. In Latin America, Brazil's LGPD gives this model an operational backbone through the ANPD, which gained the power to impose administrative sanctions once its dosimetry regulation took effect in 2023 and has since exercised that power across firms of different sizes. Chile's Law 21.719 lays out a comparably strong rights structure, among them an Article 8 bis protection against decisions based solely on automated processing, with guarantees of human intervention, explanation, and review; that law enters into force only on 1 December 2026, and its data-protection agency is still being assembled.

Our main contributions are:

- 1. An auditable instrument: a fourteen-requirement rubric anchored to the EU AI Act and the GDPR, implemented as a retrieval-grounded LLM judge that pairs every verdict with a literal evidence snippet and reaches 67.9% agreement with expert annotation on the LGPD and Law 21.719.*
- 2. Parity probe: the same instrument applied to the privacy policies of OpenAI and Anthropic across the EU, Brazil, and Chile, yielding verifiable textual evidence of a gradient in declared safeguards.*

3. *Safety-relevant finding: evidence that provider localization tracks operational institutional capacity more closely than the ambition of the legal text, since Brazil earns an ANPD-facing annex from one provider and Chile, with a substantively comparable law, stays in a generic tier.*

2. Related Work

Prior work explains how European rules travel across borders, and two strands frame our approach. The BILETA access-request study traces how data-subject rights and territorial scope shape what users beyond the EU actually receive when they exercise those rights, while a comparative-law matrix in the Global Journal of Comparative Law scores the de jure spread of GDPR-style obligations across non-EU laws by component, an approach our fourteen-requirement instrument extends to the AI Act and the GDPR in Latin America. Moving from law on the books to behaviour in markets, platform-strategy research shows that the regulation reshaped tracking practices and market structure well beyond the EU, yet it also finds that de facto adoption of the strict standard hinges on destination conditions such as regulatory capacity and credible enforcement, the very conditions our ANPD-versus-Chile contrast brings into view. Where other literature examines whether or not a national law reflects the European standard, our contribution shifts the focus to provider behaviour by measuring what specific companies declare to users in each locale.

This focus on extraction and asymmetry draws an interpretive frame from work on data colonialism, in the sense developed by Couldry and Mejias, which explains how weaker safeguards map onto broader structures of unequal data extraction and why folding Latin American users into a generic Rest of World category matters beyond doctrinal comparison, reframing uneven provider safeguards as a structural distribution of risk and not merely a compliance question. The method that makes such a reframing testable also has precedents, since privacy policies can be turned into structured evidence through machine-assisted parsing and requirement mapping, and recent compliance-classification benchmarks reporting accuracy between roughly 53% and 75% across tasks and jurisdictions situate our own result within that range. Domain-specific retrieval studies across regulated fields close the loop, showing that careful retrieval and context design improve faithfulness and reduce hallucination, which is what makes a retrieval-grounded judge dependable for legal and policy text once every verdict carries a literal evidence snippet a reviewer can verify. What no prior approach yet offers is a reusable, auditable instrument connecting the strength of a Latin American law to the safeguards a provider actually declares to users in that jurisdiction.

3. Methods

Framed within design science, the project distinguishes a design problem that builds an artefact from a knowledge question that examines part of the world. At the design level, it addresses the audit of regulatory gaps in Latin America through a retrieval-grounded judge that reproduces expert legal annotation, so that regulators and civil society can locate where provider safeguards diverge from local law. The knowledge question is descriptive. It asks whether the protection a provider declares toward Latin American users varies with the strength of the local law or remains consistent across the region. This is treated as a question of variation, since a comparison of policy texts can establish association between legal context and declared protection and cannot by itself establish cause. Within Wieringa's engineering cycle, the work covers problem investigation, treatment design, and treatment validation. Implementation in a real institutional setting and evaluation in use remain future work. In this design, the scoring engine serves as the artefact, and the parity probe reuses that engine as a measurement instrument.

Declarative compliance parity is the central construct, referring to the protection a service states in its published policy and not the technical or security parity of what the service does inside its systems. Defined this way, a missing declaration indicates that a safeguard is not promised, which already carries weight under a transparency regime such as the LGPD, even as that absence stops short of proving that the underlying practice is itself missing.

Against that construct, Brazil serves as the reference jurisdiction for the regional comparison, since the LGPD is in force, the ANPD applies administrative sanctions under its 2023 dosimetry regulation, and its investigations and formal recommendations against generative-AI providers reveal a framework that already exerts pressure, whereas Chile enters as an emerging case that stays outside the core comparison, because Law 21.719 defers its force to 1 December 2026 with its agency still being assembled.

The instrument is a fourteen-requirement rubric stored in rubrica.json. Each requirement includes an identifier from R01 to R14, a dimension drawn from five categories, a layer marker for AI Act or GDPR origin, the reference article, an es_sonda flag, and expert ground truth for each law with verdict, article, and rationale. Six requirements are sondable, meaning a provider could plausibly state them in a privacy policy. The remaining eight are institutional obligations that live in legislation.

At the core of both pipelines sits a shared judge that takes a source text and one requirement and returns structured JSON, giving a verdict, the governing article or section, a literal evidence snippet of up to fifteen words, a confidence score, a short rationale, and a human-review flag. That snippet is the central auditability choice, since any reviewer can confirm a verdict through a simple text search in the source. Feeding this judge, multilingual embeddings from paraphrase-multilingual-mpnet-base-v2 handle a corpus that mixes Spanish, Portuguese, and English, retrieving for each requirement the seven nearest chunks before the model is consulted, a value calibrated against ground truth after lower settings dropped legally relevant articles in critical cases. The two pipelines then apply this machinery to different sources. The first, motor.py,

loads the LGPD and Law 21.719 and segments them by article through a rule that splits only on article headings to avoid false cuts on internal references, after which candidate passages are retrieved for each requirement and every verdict is compared against expert annotation across 28 cells to produce scoring.json. The second, sonda.py, handles OpenAI through separate jurisdictional files and Anthropic through one policy with regional annexes, building for Anthropic logical variants for the EU, Brazil, and Chile that combine general sections with the relevant annex, and then querying the model for each provider, locale, and requirement with an explicit statement of the user's jurisdiction, a step that prevents false positives because a mention of OpenAI Ireland or Anthropic Ireland could otherwise read as coverage for a Chilean user. This second pipeline yields sonda.json across 36 cells, six requirements by three locales by two services, and validation rests on the gold set for the engine and manual verification of the probe cells.

4. Results

4.1 Regulatory scoring engine

Matching expert ground truth in 19 of 28 cells, the engine reaches 67.9% overall, breaking down to 64.3% for Law 21.719 (9 of 14) and 71.4% for the LGPD (10 of 14), with each verdict and its agreement plotted across the fourteen requirements in Figure 1. The nine mismatches are not scattered noise, since they cluster around doctrinally sensitive points such as the scope of automated-decision rights, the strength of the human-intervention guarantee, and the treatment of biometric data, and in every case the disagreement reflects a difference in legal threshold more than a misreading of source text. The clearest instance is the LGPD automated-decision requirement, where the engine correctly identifies Article 20 as allowing a request for review without guaranteeing prior human intervention and therefore marks the requirement partial where the expert records it absent, a pattern that aligns with known behaviour in retrieval-grounded judging for specialized domains, in which a model reads text correctly and still misses a stricter threshold a human expert applies, producing a doctrinal disagreement instead of a hallucination. Counting the cases in which the rationale identifies both the right article and the right logic lifts text understanding to about 85.7%, so we treat the engine's agreement rate as the answer to the requirements-satisfaction question and the 85.7% as evidence that the artefact is usable as an assistive, auditable instrument, even as a real deployment would demand higher accuracy. That score also lands within the 53–75% range of recent benchmarks while carrying two harder burdens at once, a double standard spanning the GDPR and the AI Act, and a multilingual corpus.

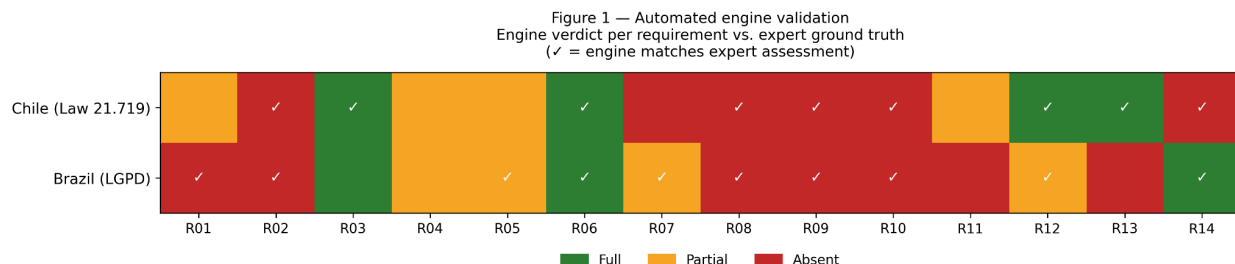


Figure 1. Engine validation against expert ground truth. Each cell gives the retrieval-grounded judge's verdict for one of the fourteen requirements, R01 to R14, under Chile's Law 21.719 and Brazil's LGPD, with full shown in green, partial in amber, and absent in red. A check mark marks agreement with the expert annotation, so the nine unchecked cells are the disagreements, leaving the engine in agreement on 19 of 28 cells, or 67.9%. Most of those mismatches fall on doctrinally contested requirements such as the scope of automated-decision rights and the treatment of biometric data.

4.2 OpenAI and Latin America

Applying the validated engine as a measurement instrument, we turn to the descriptive question of whether declared protection varies with the strength of local law, and only two of the six sondable requirements produce informative variation, namely R05 on explicability and R06 on international transfers, mapped in the parity matrix of Figure 2. On R05, the European policy is marked partial because it points users to model and training documentation, whereas Brazil and Chile receive absent, since their localized variants offer no comparable language on the logic and consequences of automated processing. The same shape returns on R06, where the June 2026 European policy earns full for citing GDPR Article 45 adequacy decisions and Standard Contractual Clauses under Article 46(2)(c), while Brazil and Chile fall to partial because the April 2026 Rest of World policy states only that transfers rely on legally valid mechanisms, naming neither a safeguard nor a local authority. Read together, these cells trace a consistent pattern, with Europe receiving the strongest package and the two Latin American jurisdictions sharing a non-EU tier even where Brazil already holds a mature framework and an active authority. That generic tier also omits the per-purpose legal-bases table present in the European policy, along with the explicit right to object to processing based on legitimate interests, the very basis OpenAI states for model training inside the EU.

4.3 Anthropic and jurisdiction

Anthropic publishes a single policy with regional annexes, and while its Brazilian annex cites the LGPD and the ANPD, no equivalent exists for Chile, a structural asymmetry that the requirements then make visible. On R03, covering automated decisions, the policy carries a global clause that the judge classifies as full across the EU, Brazil, and Chile, establishing a baseline parity that holds for every locale on that safeguard. The informative divergence surfaces on R06, where the EU and

Brazil both reach full, Brazil through the annex that cites LGPD-facing safeguards and contractual clauses approved by the ANPD, and Chile drops to absent because the transfer guarantees are framed for EEA, UK, and Swiss users alone and never localized for it. That contrast yields a clear gradient in which the EU equals Brazil and both exceed Chile, which in turn demonstrates that localized behavioural adaptation to Latin American institutional capacity is operationally possible and, in Brazil's case, already in place.

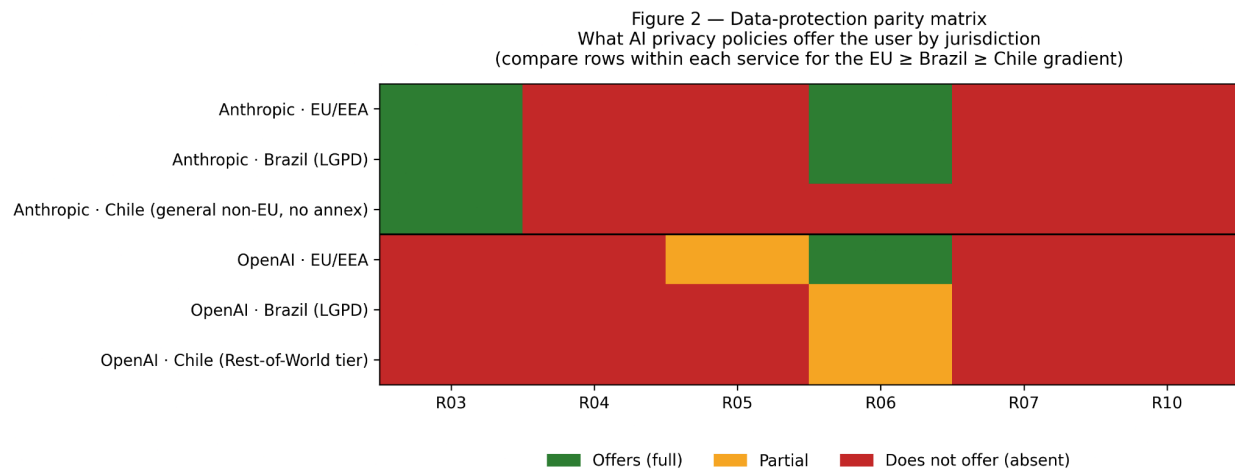


Figure 2. Declarative parity across jurisdictions. Each cell indicates what a provider's published privacy policy offers a user in a given jurisdiction across the six requirements the instrument could probe, with offers in green, partial in amber, and does not offer in red. Comparing the three rows within each provider reveals the gradient by jurisdiction. Anthropic gives Brazil a dedicated LGPD annex and leaves Chile under the general non-EU regime with no localized annex, so the EU and Brazil rows match while Chile drops, whereas OpenAI groups Brazil and Chile in a single Rest of World tier below the EU. Every verdict is anchored to a literal snippet from the source policy.

4.4 Interpretation of the parity finding

The two cases support a narrow inference, that provider safeguards track institutional salience more closely than legal ambition, a reading treated here as an abductive explanation of the gradient rather than a proven cause. The pattern is visible in the contrast itself, since Brazil's active ANPD earns differentiated recognition in Anthropic's policy while Chile's Law 21.719, for all its detailed automated-decision protections, stays folded into a generic category across both providers. That inference rests on direct quotation more than on aggregate statistics, because every cell is anchored to a verbatim snippet a reviewer can locate in the source, and with so few cells we make no claim of statistical significance.

Outside of Latin America, a recent case illustrates this same enforcement aspect. When the Korean Personal Information Protection Commission sanctioned OpenAI in 2023 for failing to report a

data leak affecting ChatGPT Plus users, it used the same decision to flag the absence of a Korean-language policy and a domestic representative, and OpenAI responded by introducing a Korea-specific privacy annex and naming a local representative, aligning its structure with what the authority had identified even though the monetary fine itself was small. That annex now sits alongside the EU-specific policy while Brazil and Chile remain in a generic Rest of World tier, a divergence that would not exist if legal text alone drove behaviour, since the two would already show similar adaptation. What better tracks localization, then, is the combination of institutional visibility and credible enforcement.

5. Discussion and Limitations

The results carry a direct implication for AI safety, since a user's ability to understand and contest harm depends on transfers, automated-decision pathways, and access to explanation, so when those safeguards are named for a European user and withheld from a Chilean one for the same service, parity of declared protection becomes part of the safety architecture itself. The variation we observe tracks institutional conditions more than doctrinal ambition, and the interpretation in Section 4.4 already traces that gradient through the Brazil and Chile contrast and the Korean parallel; what matters here is the implication, that institutional capacity and credible enforcement, rather than statutory text alone, govern how safeguards travel across borders. One boundary keeps the reading honest, that the instrument measures declared protection and not what the running system does, so the finding maps where accountability has been made legible without proving discriminatory processing.

Limitations

The load-bearing assumption is that what a policy says proxies what a user can claim, so a missing declaration marks an unkept promise more than a proven absence of practice. Were a provider to honour a safeguard it never publishes, the gradient would overstate the gap in lived protection and still describe the gap in transparency that the LGPD governs. The base is narrow, six sondable requirements and only two yielding informative variation, supporting a qualitative, text-citable reading and not statistical generalization. The judge raises its own concern, since Claude evaluated both pipelines including Anthropic's own policy, so the visible failure mode is self-favouring judgement, which we counter by anchoring every verdict to a literal snippet a reviewer can confirm, and dependence on one model still remains unresolved. A failure mode we did not address is retrieval miss, where the seven-chunk window omits a governing article and biases a verdict toward absent. Also, ground-truth labels encode doctrinal positions, so a different interpretation of GDPR Article 22 would move a few cells, and scope covers only Brazil, Chile, OpenAI, and Anthropic.

Future Work

Next steps follow from those limits, extending the instrument to Colombia and Mexico, adding providers and safeguards, and moving from declarative parity toward a technical audit of system behaviour. A longitudinal run offers the cleanest test of the causal claim, tracking whether providers localize once a national authority is operational, with Chile's 2026 activation serving as a near-term natural experiment.

6. Conclusion

A method that makes a safeguard either present or absent in a provider's own words changes what accountability can rest on. When a commitment must be located as a literal line in a published policy, the burden shifts away from trusting that a global rule diffuses on its own and toward asking what a specific company has chosen to write down for a specific reader. That move is modest in mechanism and consequential in framing, because it treats the privacy policy not as boilerplate but as an instrument of differential treatment that can be read, compared, and held to account. The same lens that surfaces a gap for one region can be turned on any provider and any jurisdiction, which is the property that makes an auditable comparison worth more than a single finding.

The deeper reflection is that protection on paper and protection in practice come apart precisely where institutions are weakest, and this is the place where users have the least capacity to notice or repair the difference on their own. A right that no provider translates into a localized commitment asks the individual to know the law better than the company that processes their data, an expectation that inverts where responsibility should sit. Treating public-facing parity as a safety concern reframes that inversion as something measurable and therefore contestable, and it suggests that the work of equal protection in cross-border AI is not finished when a strong law is written. It continues through the slower labor of building the institutional presence that makes a law worth localizing, and tools that render the gap visible are one small part of holding that process to its promises.

Code and Data

Include links if applicable. If your project doesn't involve code (e.g., policy analysis) or if there are info-hazard considerations, note that here.

- **Code repository:** <https://github.com/PMeik/paridad->
- **Data/Datasets:**

Ley 21.719 (Chile): [<https://www.bcn.cl/leychile/navegar?idNorma=1209272>]

LEI N^o 13.709 (Brasil):

[https://www.planalto.gov.br/ccivil_03/ato2015-2018/2018/lei/l13709.htm]

Openai-eu: [<https://openai.com/es-419/policies/eu-privacy-policy/>]

Openai-row: [<https://openai.com/es-419/policies/privacy-policy/>]

Anthropic: [<https://www.anthropic.com/legal/privacy/>]

References

1. Alabbasi, N., Erak, O., Alhussein, O., Lotfi, I., Muhaidat, S., & Debbah, M. (2024). *TeleOracle: Fine-Tuned Retrieval-Augmented Generation with Long-Context Support for Networks*. arXiv:2411.02617. <https://arxiv.org/abs/2411.02617>
2. Anthropic. (2026). Privacy Policy. <https://www.anthropic.com/legal/privacy>
3. Autoridade Nacional de Proteção de Dados (ANPD). (2023). Resolução CD/ANPD nº 4, de 24 de fevereiro de 2023: Regulamento de Dosimetria e Aplicação de Sanções Administrativas. Diário Oficial da União, Brasil.
4. Brasil. (2018). Lei nº 13.709, de 14 de agosto de 2018 (Lei Geral de Proteção de Dados Pessoais, LGPD). https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm
5. Chile. (2024). Ley N° 21.719 que regula la protección y el tratamiento de los datos personales y crea la Agencia de Protección de Datos Personales. Diario Oficial, 13 de diciembre de 2024. <https://www.bcn.cl/leychile/navegar?idNorma=1209272>
6. Couldry, N., & Mejias, U. A. (2019). Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject. *Television & New Media*, 20(4), 336–349. <https://doi.org/10.1177/1527476418796632>
7. European Parliament and Council. (2016). Regulation (EU) 2016/679 (General Data Protection Regulation, GDPR). Official Journal of the European Union. <https://eur-lex.europa.eu/eli/reg/2016/679/oj>
8. European Parliament and Council. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
9. Harkous, H., Fawaz, K., Lebre, R., Schaub, F., Shin, K. G., & Aberer, K. (2018). Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning. 27th USENIX Security Symposium, 531–548. <https://www.usenix.org/conference/usenixsecurity18/presentation/harkous>
10. OpenAI. (2026). EU Privacy Policy. <https://openai.com/es-419/policies/eu-privacy-policy/>
11. OpenAI. (2026). Privacy Policy (Rest of World). <https://openai.com/es-419/policies/privacy-policy/>
12. Personal Information Protection Commission (PIPC). (2023, July 26). Decision imposing an administrative fine on OpenAI OpCo, LLC for failure to report a data breach. Republic of Korea.
13. Ran, H., Zhang, H., & Tang, X. (2025). GDPR-Bench-Android: A Benchmark for Evaluating Automated GDPR Compliance Detection in Android. arXiv:2511.00619. <https://arxiv.org/abs/2511.00619>

14. *Safety Compliance: Rethinking LLM Safety Reasoning through the Lens of Compliance (Compliance Reasoner)*. (2025). *arXiv:2509.22250*. <https://arxiv.org/abs/2509.22250>
15. Wieringa, R. J. (2014). *Design Science Methodology for Information Systems and Software Engineering*. Springer. <https://doi.org/10.1007/978-3-662-43839-8>
16. Wilson, S., Schaub, F., Dara, A. A., et al. (2016). *The Creation and Analysis of a Website Privacy Policy Corpus (OPP-115)*. *Proceedings of the 54th Annual Meeting of the ACL*, 1330–1340. <https://doi.org/10.18653/v1/P16-1126>
17. Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2025). *Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools*. *Journal of Empirical Legal Studies*, 22(2), 216–242. <https://doi.org/10.1111/jels.12413>
18. Mahieu, R., Asghari, H., Parsons, C., van Hoboken, J., Crete-Nishihata, M., Hiltz, A., & Anstis, S. (2021). *Measuring the Brussels Effect through Access Requests: Has the European General Data Protection Regulation Influenced the Privacy Rights of Canadian Citizens?* Working paper presented at BILETA 2021, April 14, 2021.
19. Eshkita, R., & Stamhuis, E. (2024). *The Influence of the Brussels Effect on the Interpretation of Data Protection Laws in the Gulf*. *Global Journal of Comparative Law*, 13(2), 261–278.
20. Peukert, C., Bechtold, S., Batikas, M., & Kretschmer, T. (2020). *European Privacy Law and Global Markets for Data*.
21. Kresevic, S., Giuffrè, M., Ajcevic, M., Accardo, A., Crocè, L. S., & Shung, D. L. (2024). *Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework*. *npj Digital Medicine*, 7.
22. Knollmeyer, S., Caymazer, O., & Grossmann, D. (2025). *Document GraphRAG: Knowledge Graph Enhanced Retrieval Augmented Generation for Document Question Answering Within the Manufacturing Domain*. *Electronics*, 14(11), 2102.
23. Albrecht, Jan Philipp. “How the GDPR Will Change the World.” *European Data Protection Law. Review* 2, no. 3 (2016): 287-89. <https://doi.org/10.21552/EDPL/2016/3/4>

LLM Usage Statement

In this project, we use an LLM (claude-haiku-4-5, temperature 0) as a classification judge within our RAG pipeline, evaluating each requirement against legal texts and privacy policies. This is a documented and reproducible part of the methodology, since every ruling is anchored to a verbatim fragment that we independently verify. In addition to its use as a classification judge, we have utilized Claude to assist us in generating possible architectures and debugging python. Our legal documentation was reviewed by a subject matter expert and not by an LLM. All of our accuracy numbers and parity conclusions are confirmed by reviewing the output files scoring.json and sonda.json.